Unpacking the Factors Influencing Election Outcomes: A Comprehensive Analysis

Kecheng Ye

Introduction

• Elections are the foundation of any democratic system, and their outcomes have a significant impact on the governance and future trajectory of a nation.

• Concerns and issues regarding elections have emerged over the years, including voter suppression, election security, and money in politics

• It's important to understand what factors contribute to election results

• Factors that can influence election results include income, age, food, race, marriage, education, and votes & parties

• Analyzing the interplay between these variables can provide insight into predicting election outcomes

Analysis & Methodology

- Dataset Gathered:
- US Census and Election Results (2000-2020) from Kaggle. Fundamental variables include individuals' annual income, annual total family income, age, gender, marital status, race, citizenship status, language spoken at home, education level, and employment status at the individual level.

- Dataset Cleaned:
- Dataset columns are renamed such that reader will be easier to understand the features. Missing values are dropped due to vast time range, it will not be feasible to replace them.

Statistical Models:

Uncover patterns and relationships within the data, make predictions about various outcomes, and ultimately gain a deeper understanding.

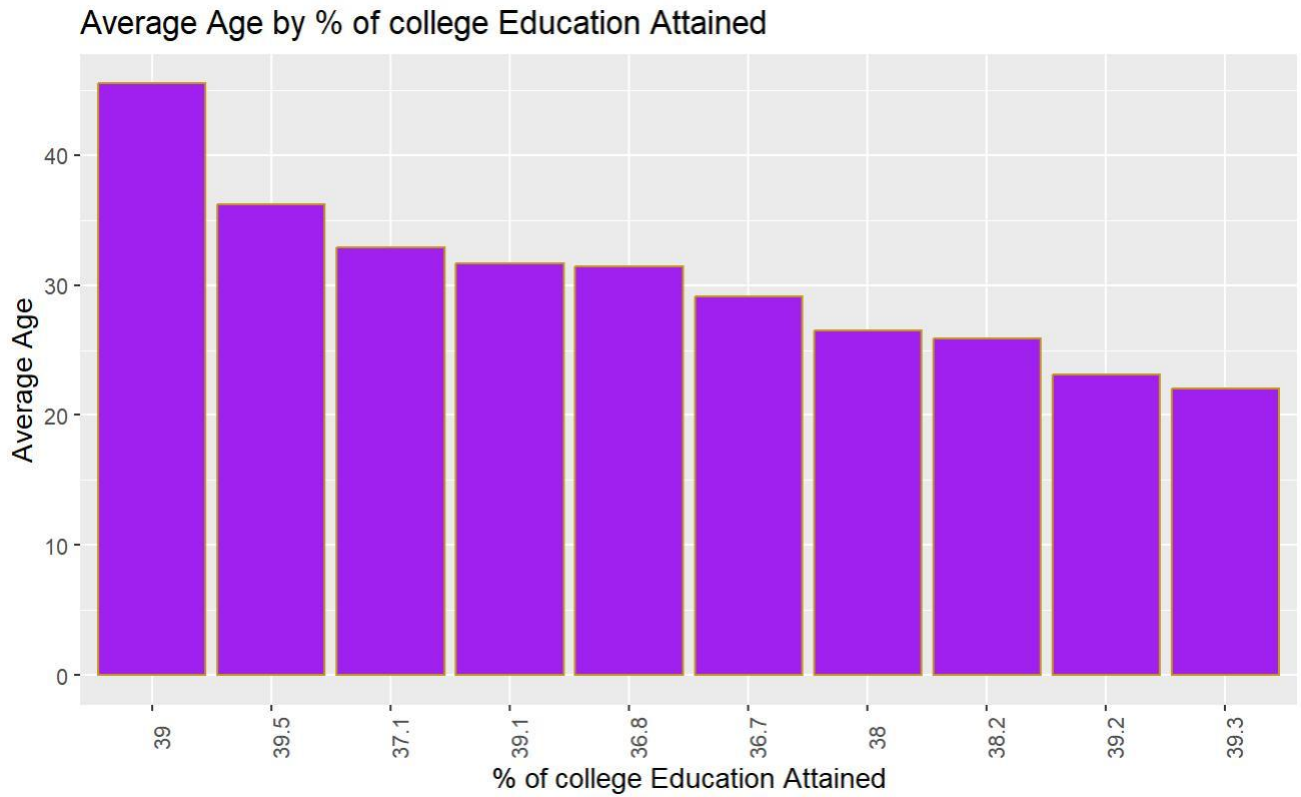
Our models:

Logistic Regression & LDA & QDA Random Forest KNN Clustering GBM & SVM XGBoost LASSO

EDA

CONTENTS

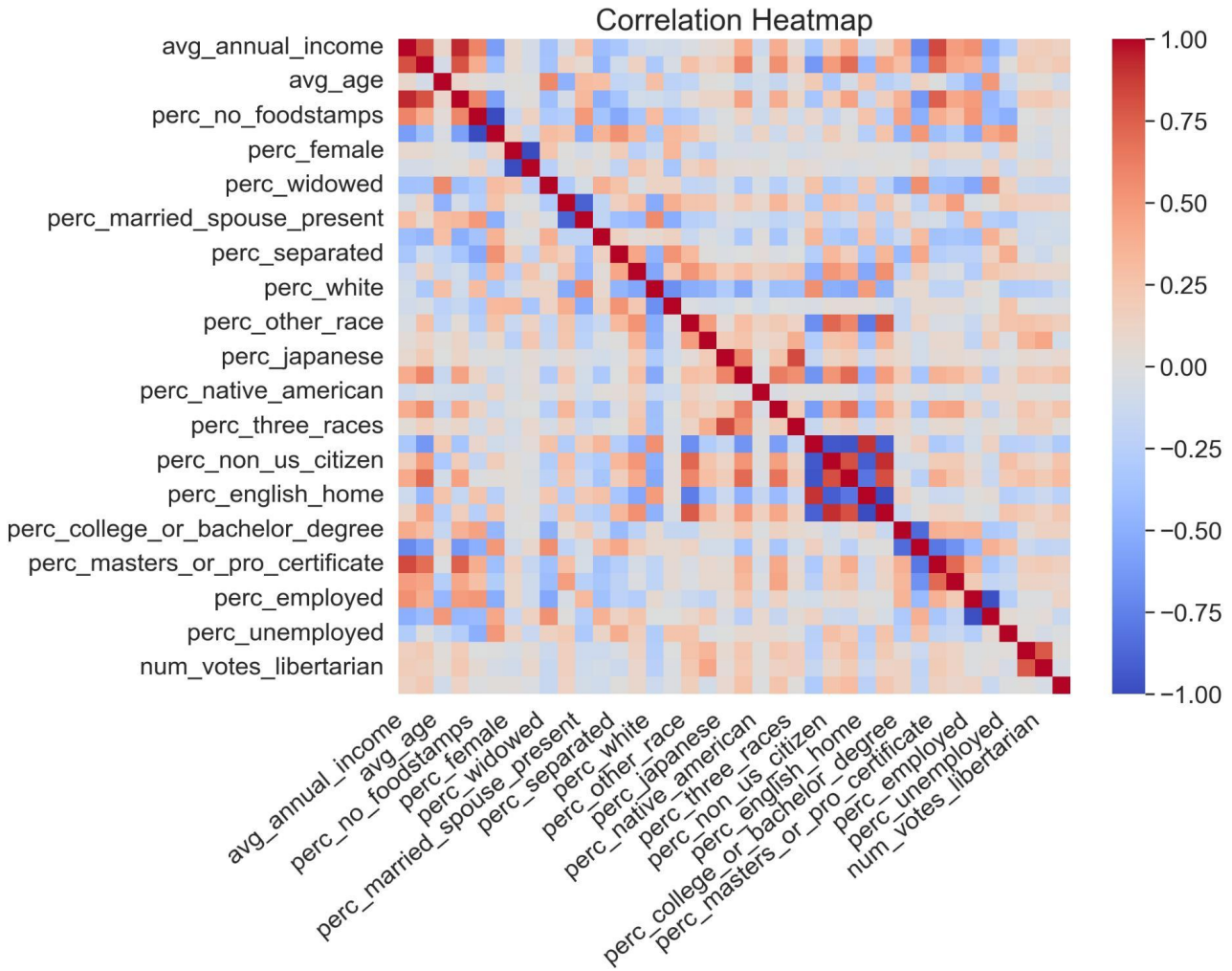
Some fundamental variables pattern and correlation: We can see that the majority electors have average age from 40 to 50.



EDA

CONTENTS

The Overall pattern and correlation among features



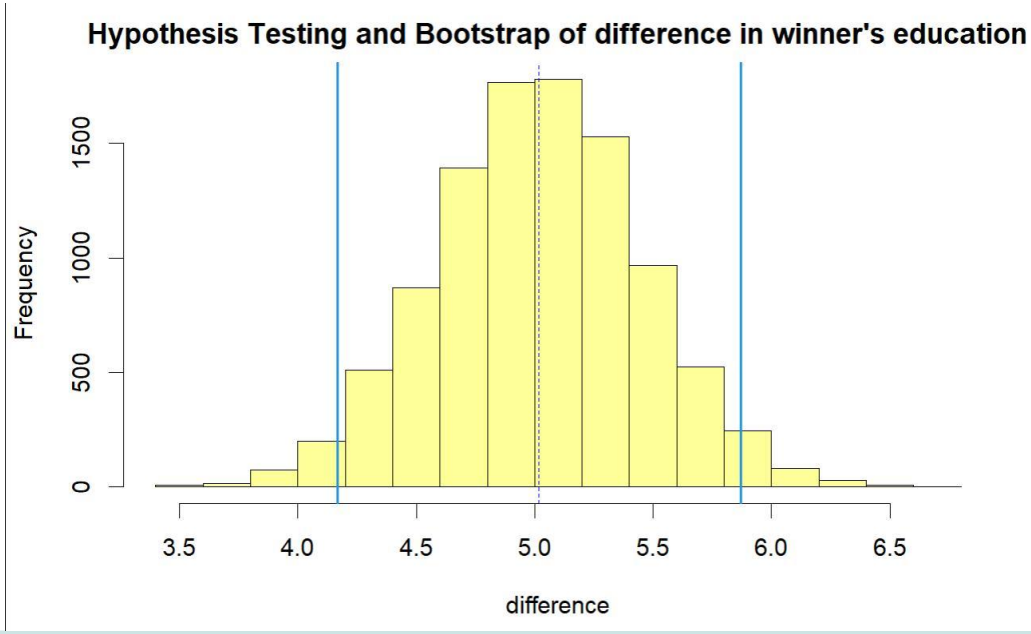
Few of the features shows strong correlation to each other. In addition, the response feature is the winner, the categorical variable.

Hypothesis Testing

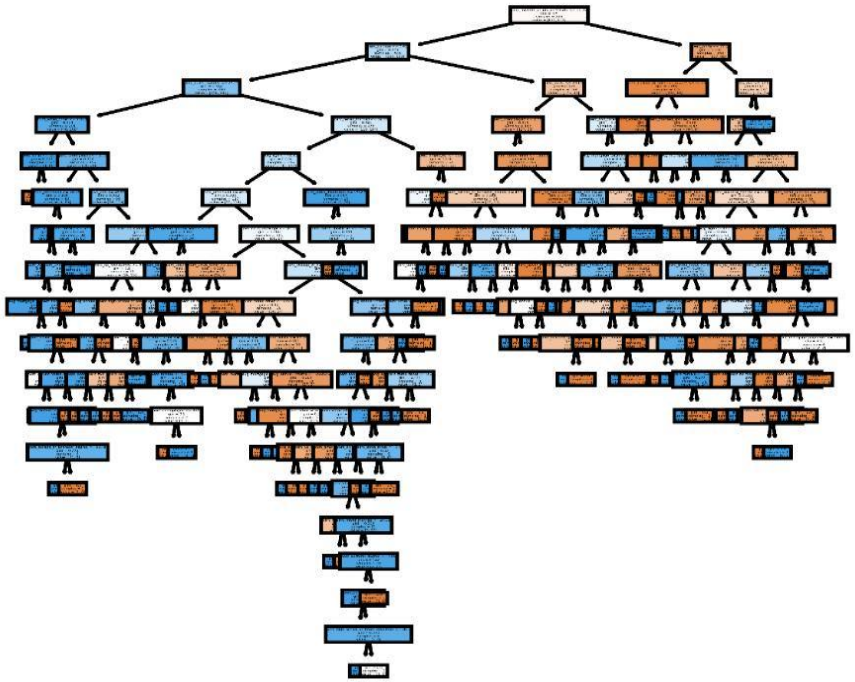
H0: The mean education for both party winners is the same

Ha: The mean education of Republican winner is higher

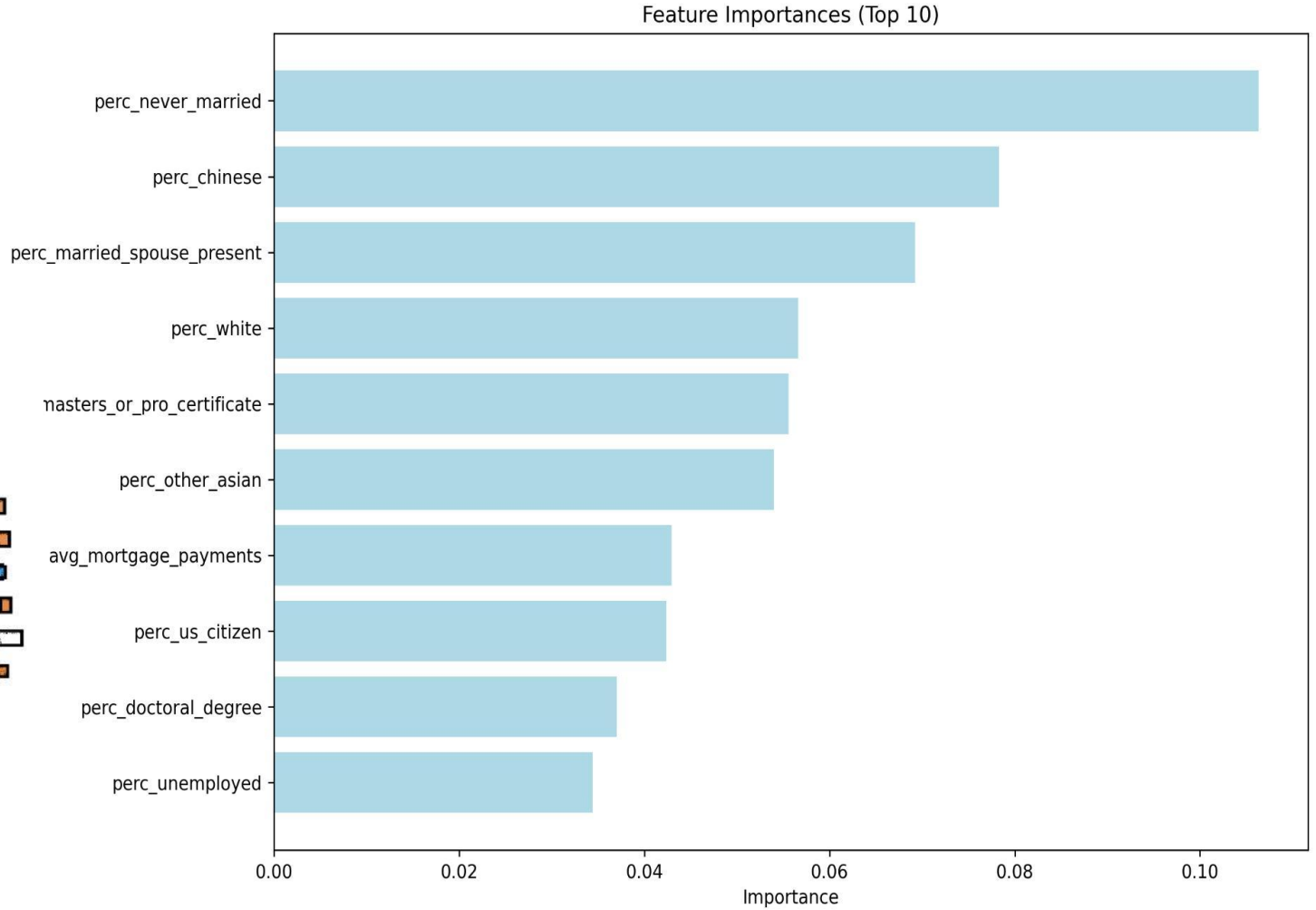
P value < 0.05



Modeling: Random Forest

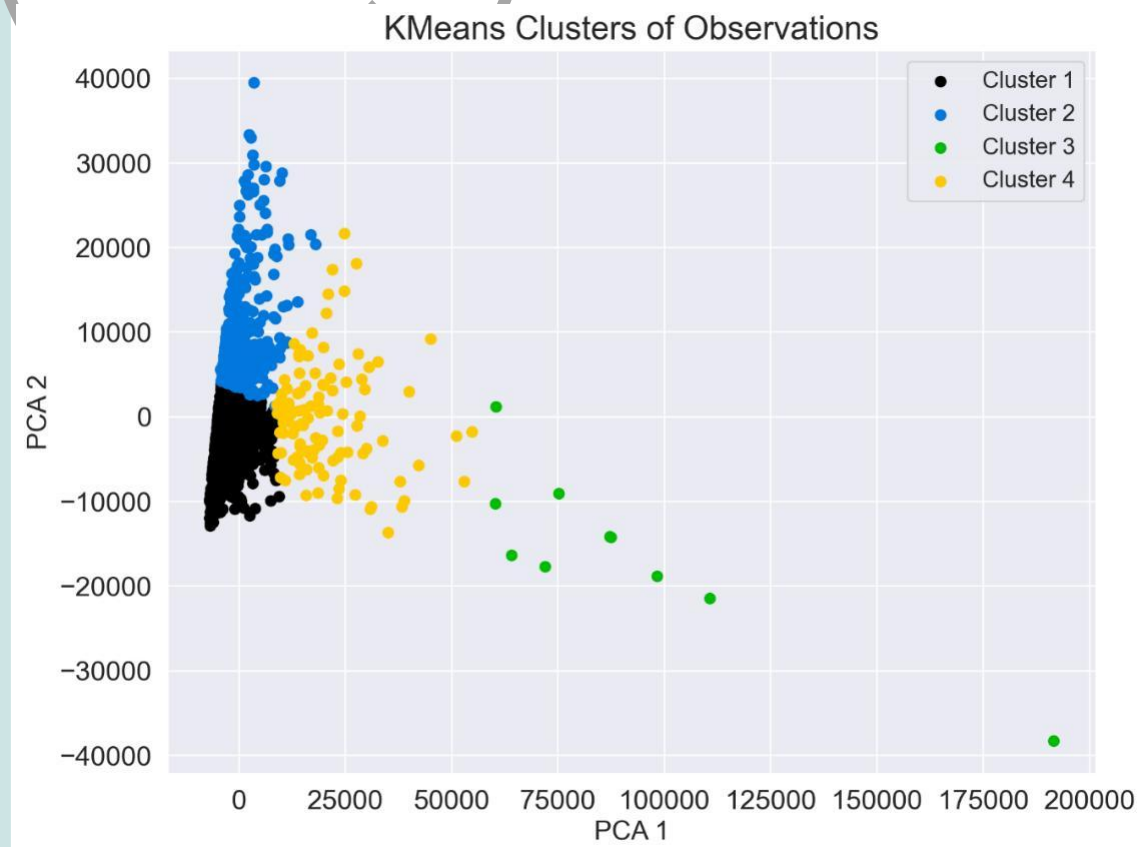


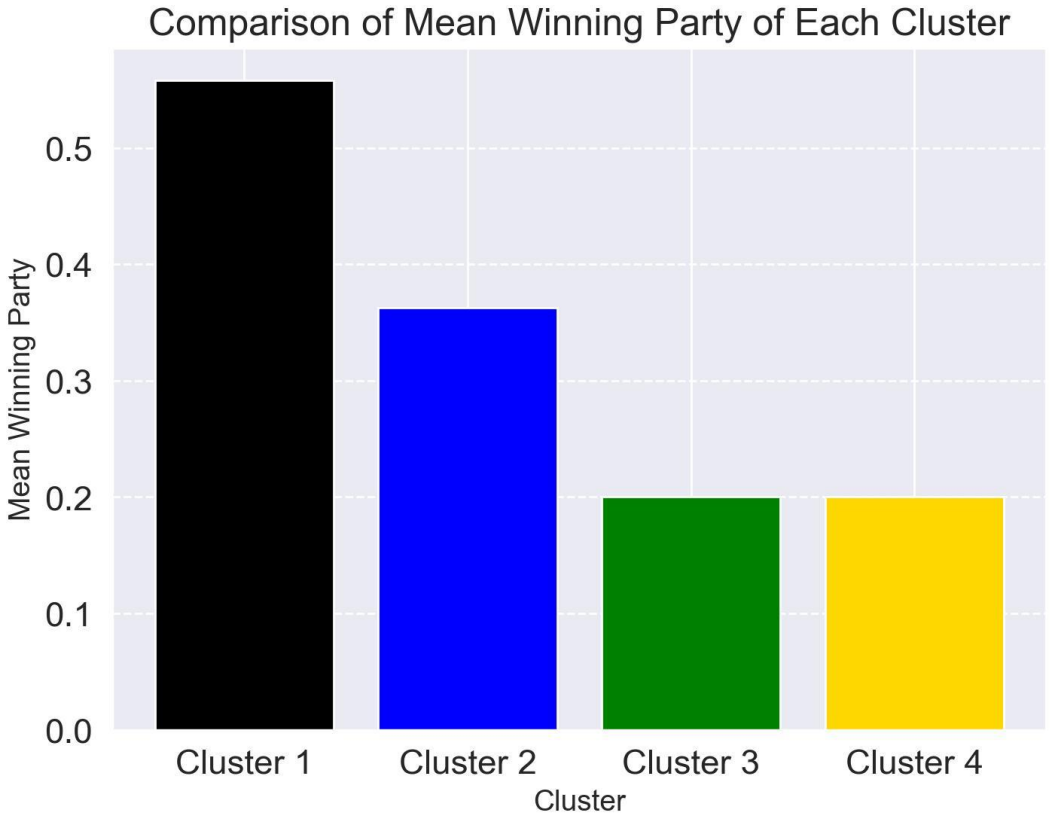
The results show that the top 10 feature that contribute mostly to election results are marriage, race, citizenship, mortgage payments, doctoral degree, and employment.



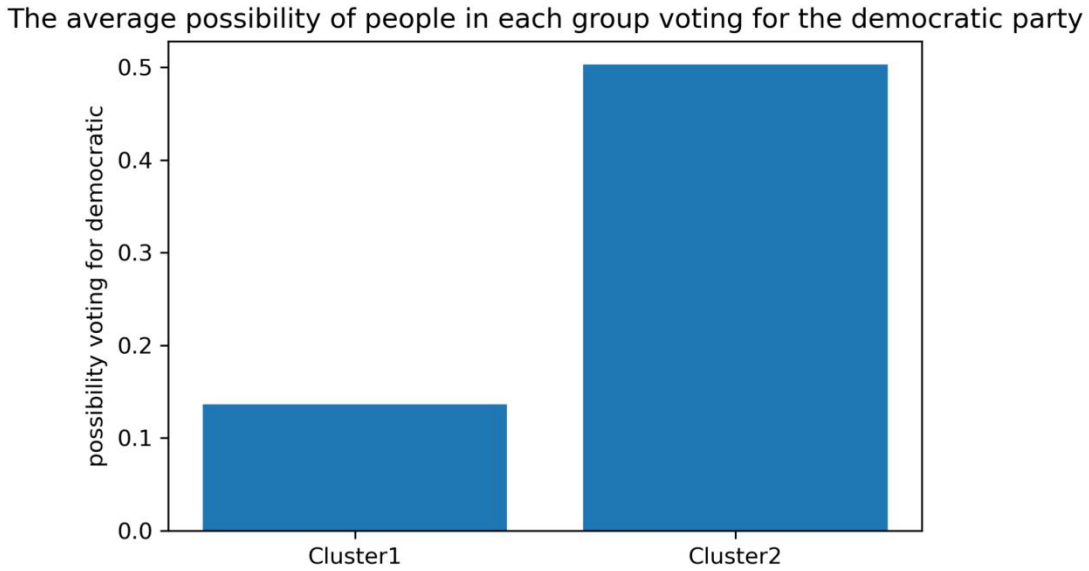
Modeling: KNN-Clustering

We can see that the cluster group 3, and 4 are balanced which cluster group 3 only contains 9 observations. • Here we can see that the cluster one group’s average winning party are significantly higher than the other three cluster group.



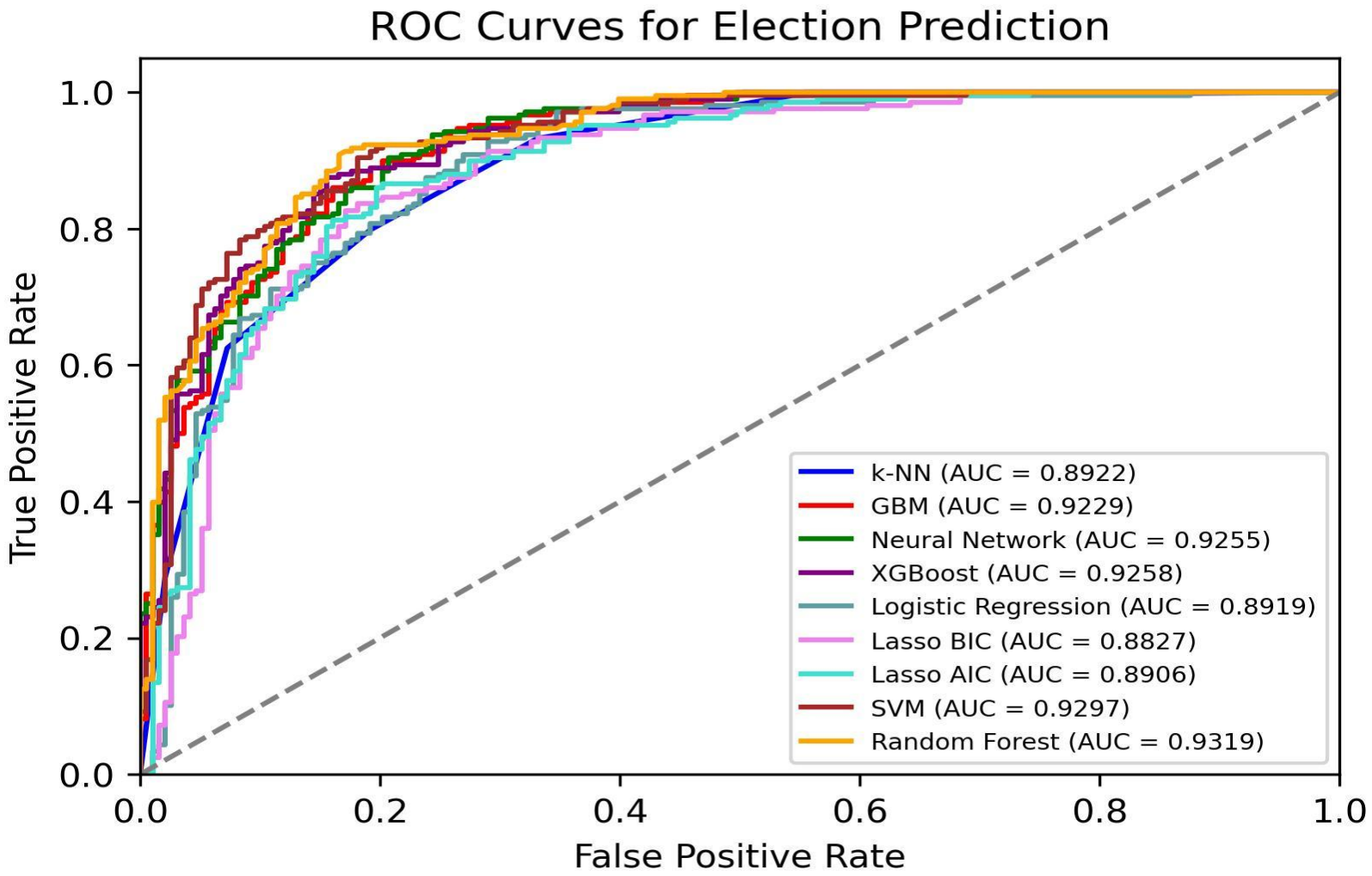


Our research team finds that people who have been put into Cluster 2 are significantly more likely to vote for Republicans compared to people who have been put into Cluster 1.



Models Comparison

We can see that Random Forest Model has the best performance in predicting factors that contribute to selection method. GBM, XGBoost models also have a relatively good performance. While KNN, logistic regression, and Lasso models has relatively low performance.



CONCLUSION

We have found that the key features that affect election outcome are actually “Never Married” & “Race Info”.

- We find that both supervised and unsupervised method are giving us a great result in term of predicting our target variable.
- There are differences among Republican and Democrat Winners: Education, Age, and some other features
- We have found that our dataset contains many features with strong correlation, and we removed them.
- Random Forest are having a better performance than XGBoost, which could because of our data set does not contains many features after data cleaning. Features such as foodstamps, family incomes, home languages do not necessarily provide a change to the election outcome.
- The combinations of different features have more to contribute to the election outcomes than single feature.